LA-UR- 04-5011

Approved for public release; distribution is unlimited.

Title:	Datamart Use for Complex Data Retrieval in an ArcIMS Application
Author(s):	Steven Scherma Stephen Bolivar
Submitted to:	ESRI International Users' Conference, San Diego, CA

LOS ALAMOS NATIONAL LABORATORY
3 9338 01000 0528

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Form 836 (8/00)

Datamart Use for Complex Data Retrieval in an ArcIMS Application

Steven Scherma, Stephen Bolivar, PhD
RRES-Environmental Characterization and Remediation
Risk Reduction and Environmental Stewardship Division
Los Alamos National Laboratory

Abstract

This paper describes the use of datamarts and data warehousing concepts to expedite retrieval and display of complex attribute data from multi-million record databases. Los Alamos National Laboratory has developed an Internet application (SMART) using ArcIMS that relies on datamarts to quickly retrieve attribute data, associated with, but not contained within GIS layers. The volume of data and the complex relationships within the transactional database made data display within ArcIMS impractical without the use of datamarts. The technical issues and solutions involved in the development are discussed.

Background: General

Los Alamos National Laboratory (LANL) was founded in 1943 as part of the Manhattan Project and subsequently developed the United States' first atomic weapon. The Environmental Restoration Project (now Remediation Services) was established in 1989 by the Department of Energy to remedy environmental problems which have occurred during the past 50 years of Laboratory operations. Remediation Services is responsible for the characterization, clean up, and monitoring of over 2,124 identified potential release sites (PRS) largely associated with past weapons research, development, and production activities. To accomplish project goals, Remediation Services conducts field sampling activities to determine possible types, levels and geographic extent of chemical contamination. These sampling activities have generated approximately 3 million analytical chemistry records which, in turn, have supported several hundred regulatory deliverables.

Background: Information Management System

The primary goals of the Remediation Services' information management system have been identified as:

1) increasing the timeliness and efficiency of regulatory deliverable production,
2) insuring deliverables are generated and maintained with standard extraction, visualization and reporting standards
3) and insuring deliverables are properly tracked and audited.

As a result of these goals, Remediation Services is currently building a system that supports a defensible regulatory document production, provides for cost-effective data management, and allows for high-quality data access and visualization. The access and visualization application that has been developed and deployed is the SMART Environmental Data Retrieval and Visualization application.

The overall Information Management system design (see Figure 1) consists of a multiple database repositories that store tabular data in both a transactional database and a data warehouse/marts, GIS

data within a Spatial Database Engine (SDE), and documents on file servers. Applications fall into two general categories: data input applications (Sample Tracking Application) and data output/decision applications (SMART and GeoPro Modeling Management System).

Currently, the transactional database "feeds" the SMART datamart. [Note: in the future it will feed an intermediate warehouse repository which will in turn feed the datamart.] While transactional databases are used for the collection, processing and quality assurance of data, they are not optimized for data retrieval. To overcome this deficiency the SMART datamart pulls data from the transactional database and configures it in a format that makes querying, retrieving and reporting on data quick, easy and efficient.

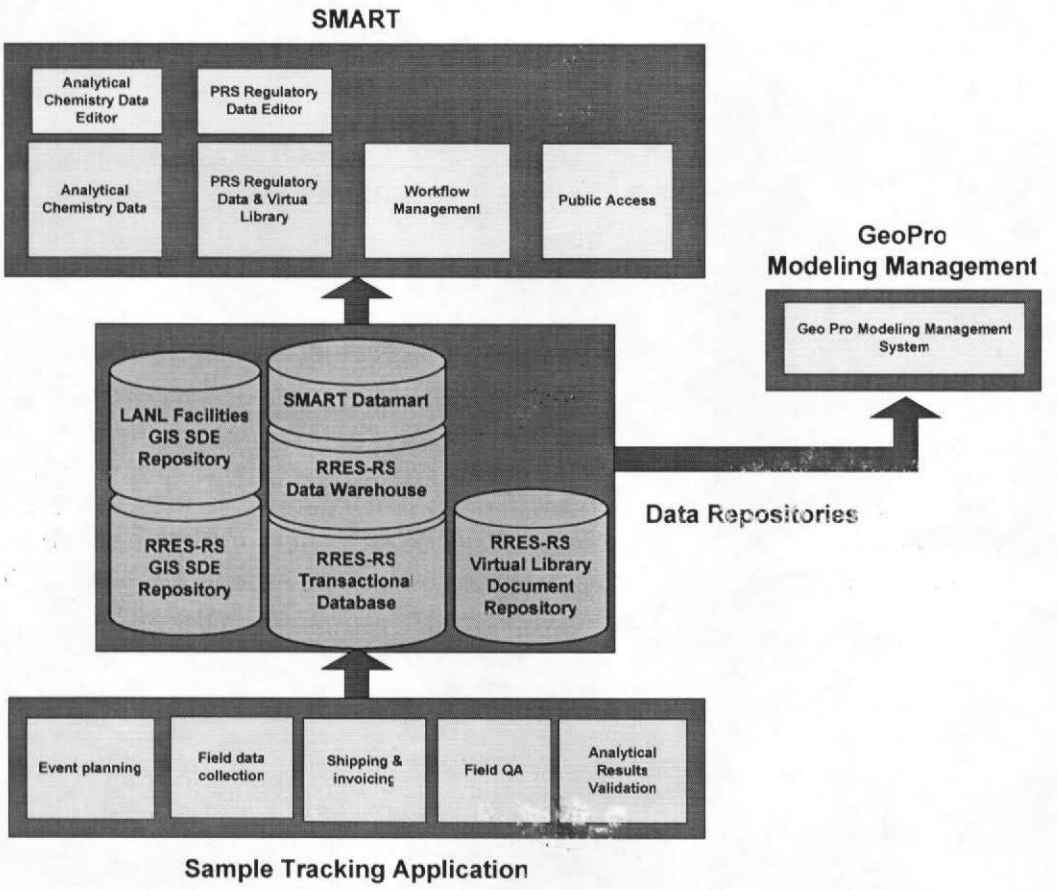


Figure 1: Information Management System Conceptual Design

Background: SMART Environmental Data Retrieval & Visualization Application

Description of SMART

SMART is a web-based data retrieval and visualization tool that allows users to draw from multiple data repositories and gain access to a wide range of the lab's environmental chemistry data, regulatory status information, graphs, photos, regulatory deliverable reports and spatial GIS data. The SMART interface allows users a choice of extracting data in one of two ways: through a series of text based screens which allow the user to "drill down" based on a number of user selected criteria, or through a GIS map interface that allows a user to zoom in to an area of interest and select data associated with a

GIS feature. Whichever option is chosen, the user may then display their selected data directly in reports or maps or view selected documents directly in the browser.

SMART allows a user access to approximately:

- 3 million analytical chemistry results records
- 50 thousand sample records
- 20 thousand sampling location records
- 2 thousand PRS regulatory tracking records
- 3 thousand associated PRS documents
- 3 thousand associated PRS photographs
- 20 layers of GIS data including orthophotography

The main goals of the SMART are as follows:

- Ease of access to analytical, regulatory, GIS and document data
- Easy to use analysis and visualization tools
- Tracking and auditing of data sets
- Automation of reports, graphs and maps

Data Retrieval

The main interfaces of the SMART application consist of a tabular interface that allows users to drill down by location, sample and result criteria and a GIS powered map that allows a user to select data or retrieve information based on a particular geographic feature. Links within the map allow the user to branch off to related data.

The application provides an intuitive and easy-to-use interface. The user interacts with custom buttons in the GIS interface and simple pick boxes within the tabular interface. For example, a user could retrieve the 23 sample records (from 50,000) containing cadmium, above background, in rock, for available for a particular investigative area in six simple clicks (see figure 2). The user may then view all the analytical chemistry data associated with those samples, view a summary report of the data, individual sample reports, or export the data to a spreadsheet if necessary.

The user can display the locations of the samples in their data set on the GIS map to better understand the spatial distribution. Within the GIS map, users can query an individual location to retrieve its sampling and result data as well as retrieve any documents or graphs associated with those samples (see figure 3). If it is determined that the user needs to add or remove locations (and their associated samples) from their data set, that can be done directly from the map. The user can select or deselect data directly from the GIS interface using custom built tools, thereby adding additional locations (and their samples) to the dataset or removing locations (and their samples) from the dataset.

With the spatial location of their samples displayed, the user can easily determine spatially related PRSs and query for the PRS's regulatory information, by selecting a PRS feature off the map and calling up the related link. This gives the user access to all the regulatory reports, correspondence, and photographs associated with the PRS (see figure 4).

Data Visualization

One of the most powerful aspects of the SMART applications is its ability to visualize data spatially. Since all sample results are tied to physical locations and associated with PRS features, being able to understand spatial relationships is crucial. In addition to being able to quickly determine where the samples are located, it allows a user to see what other sampling activity has occurred in the area and the relationship of the sampling to surrounding structures, canyons, drainages, etc. (see figure 5). The user may take a dataset and display it in reports and graphs that have been custom tailored to the needs of the project (see figure 6). These reports are designed around the tables and graphs that are used in the regulatory deliverables required by the project.

Data Set Tracking and Auditing

Once an acceptable data set is created, the data set is saved both as reference to records in the database and as a "copy" of the data. This allows a user to run a "delta report" and note any database changes to the data set at any time in the future. Metadata is automatically associated with the saved data set, describing the data set preparation process. These data sets are submitted to a quality assurance approval process. Once approval is received the data sets may be used in deliverables. The data sets as well as the entire quality assurance process are fully tracked and audited.

When the data set is used in maps and/or reports and sent to colleagues, regulators, or other stakeholders, the relevant database records, selection criteria, and data preparation method information can be quickly accessed and understood. Since SMART pulls data sets directly from a secure source, which is be monitored, audited and reviewed, reporting activity uniformly refers back to original source data. Use of this application provides Remediation Services with the ability to better control the data used in regulatory reporting and allows for the maintenance of a strong regulatory record.

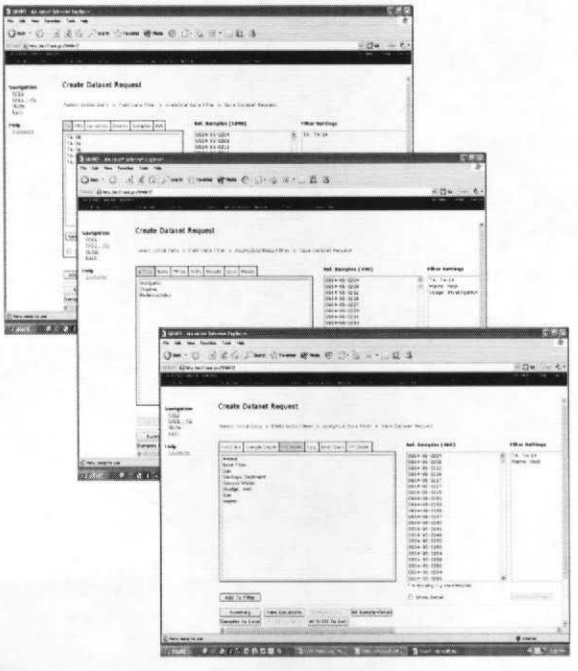


Figure 2: Tabular data retrieval in SMART

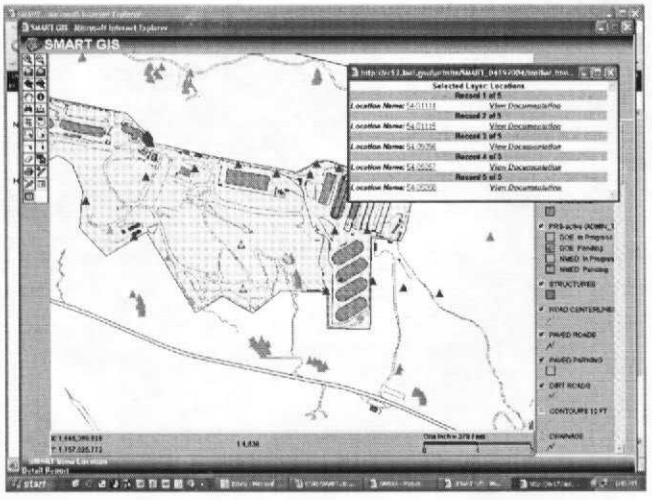


Figure 3: GIS-based data retrieval in SMART

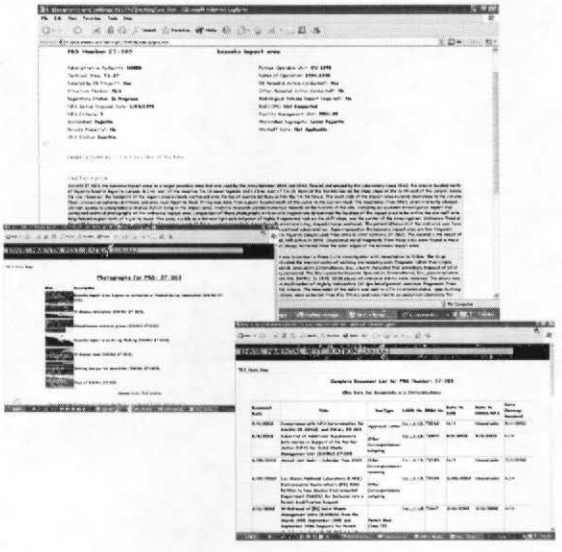


Figure 4: PRS regulatory data retrieval in SMART

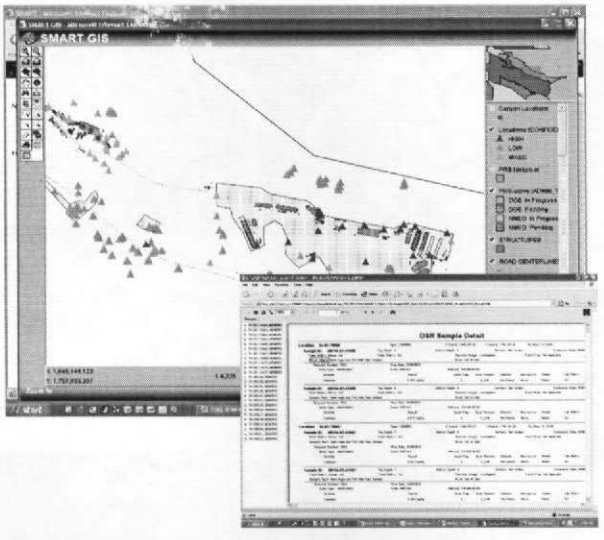


Figure 5: GIS Data Visualization in SMART

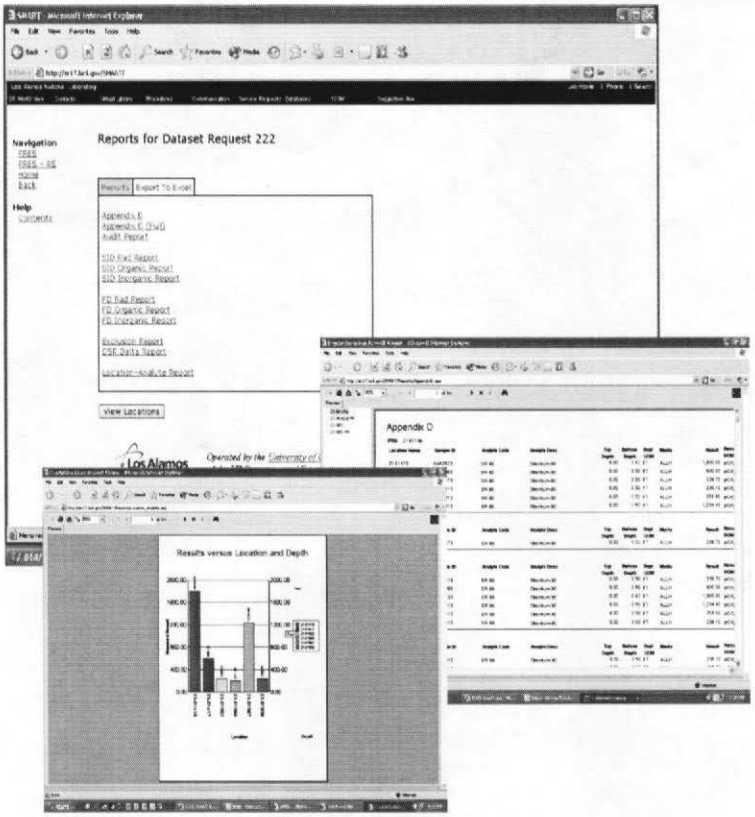


Figure 6: Tabular and graphical data visualization in SMART

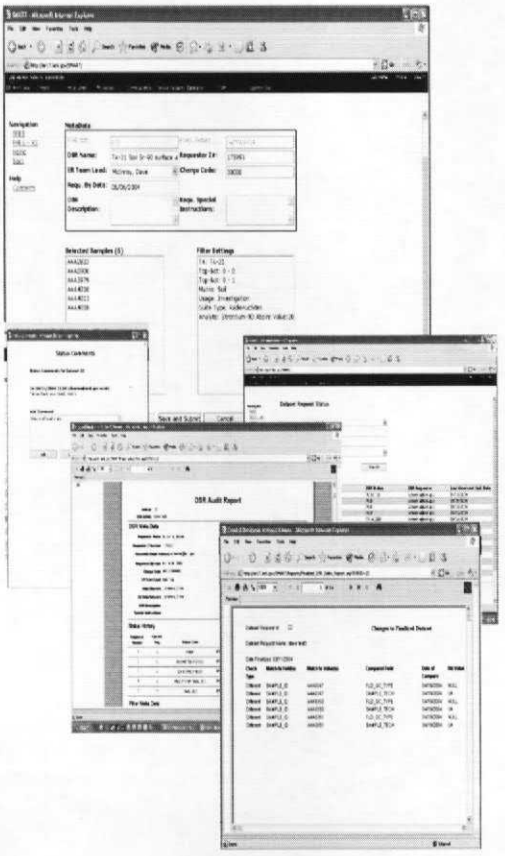
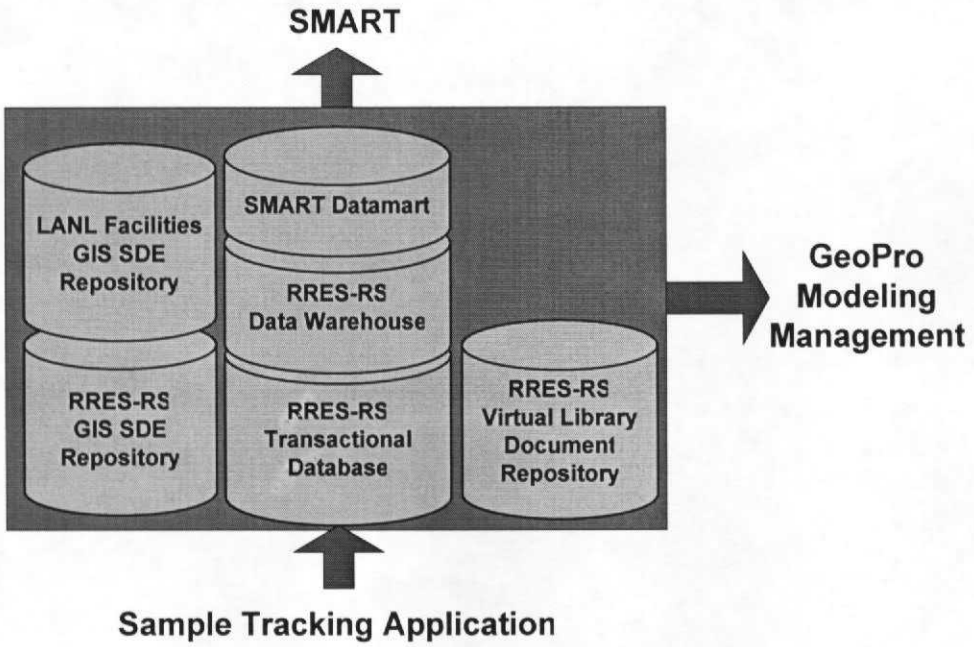


Figure 7: Tracking and auditing in SMART

Database Overview



RRES-RS Transactional Database

The RRES-RS Transactional Database is optimized to support data entry. It is a fully relational database.

RRES-RS and LANL Facilities SDE GIS Repositories

The RRES-RS GIS data resides in multiple ESRI SDE servers. RS owned data is maintained and supported under internal change control. General LANL facility data is accessed through a separate SDE server. The combined data is served on the web through an ESRI ArcIMS Server.

RRES-RS Virtual Library Document Repository

The Virtual Library Document Repository is a PDF file repository of major documents such as environmental reports produced by RS are kept and accessed on a file server.

RRES-RS Data Warehouse

The RRES-RS Data Warehouse repository will eventually be the "repository of record" for the RRES Division. Datamarts such as the SMART datamart will pull data from this repository.

SMART Data mart

The SMART Datamart for RS analytical and regulatory data is used in SMART. Indexed in a typical star schema/fact table format. This datamart is optimized for typical queries performed with SMART and for quick web display.

Database Design

The RRES-RS databases are designed with the goal of supporting all phases of RRES-RS processes. It is an ongoing, comprehensive design that allows the integration of workflow data, regulatory data, GIS data and technical data into a set of transactional and warehouse databases. The workflow data consists of project status data, management data, and business rules that support the RRES-RS workflow process. The technical data consists of field data, results and analysis data. The regulatory data consists of tracking information on regulatory decisions and the supporting document repository. The spatial data consists of the information necessary for data visualization and modeling. The resultant database is comprised of over 300 tables with a complex set of relationships as well as implied and programmed business rules.

RRES-R Transactional Database

Developing a transactional database that has and maintains data integrity, completeness and non-redundancy were primary requirements of the transactional database design. As a result, the database contains numerous checks and constraints to fulfill the design requirements.

Approximate numbers for RRES-RSDB (as of 9/15/03) are given below:

- 20 thousand location records
- 47 thousand sample records
- 3 million analytical results records
- 2 thousand PRS regulatory tracking records
- 3 thousand associated PRS documents
- 1350 database constraints
    - 650 foreign key constraints
    - 350 primary key constraints
    - 200 check constraints
    - 150 unique index constraints
- 850 data checks
    - 310 stored procedures
    - 540 auditing triggers
- 8 auditing tables
- QC tables specifically for samples and results

The transactional database supports the RRES-R data input applications. The transactional database is designed in a SQLServer relational database, but could be built in any relational database such as Oracle or DB2. The goal of the transactional database was to ensure that complete, accurate and non-redundant data is captured and stored. It was designed specifically to support transactions and is highly denormalized.

Denormalization is to attempt to capture and preserve a piece of data once and only once in the database. For example, a user name may be used in multiple tables in the database yet is only stored in one table, not in each table that contains the role. The user name is given a reference number and that number is used throughout the database when reference is needed to the name.

Look up table constraints were placed on as many fields as possible. Using look up tables requires users to pick from a list of items rather than allowing unrestrained entry. This prevents typing errors and misspellings which could lead to redundant data and difficulty in querying data. For example Steve Smith, Steven Smith and Stephen Smith may all be the same person. A query requesting data for Steve Smith would not retrieve all relevant records. Using look up tables helps reduce these types of errors.

Triggers and stored procedures help in checking a data record for completeness and errors. These are typically used to enforce business rules during the data entry process. Stored procedures can be quite complex and resemble mini software programs performing multiple checks, conversions, and the like on the data.

Primary keys, data types, field sizes, data formats, ranges and allowing null fields are additional ways to constrain data and ensure completeness and reduce redundancy in a transactional database. By utilizing these database techniques and solid design principles a transactional database can prevent or reduce a large number of human errors that can easily manifest itself in a "non-relational" database. It is generally acknowledged that designing strong transactional databases is a best practice for ensuring that accurate, complete, non-redundant data is captured by an organization.

Once a data transaction is complete, the relevant data is transferred to the data warehouse and then purged from the transactional database.

The strengths of the transactional database for data entry turn into weaknesses for data retrieval. The transactional database basically tries to atomize each piece of data. Each 'atomic" bit of information is ideally stored and managed separately to minimize error. However when data retrieval and display is required all of the disparate elements must be reassembled. This can result in enormously complex SQL statements and can produce increasingly slow results as databases grow increasingly large. For example in the RRES-R database to pull a complete analytical record may require the joining of 60+ tables. While the use of database views can help in these situations they are not optimized for reporting out data. This is where datamarts can provide benefits over a transactional database for reporting purposes.

SMART Datamart

On top of the transactional database is built the SMART Data mart. While transactional databases are used for the collection, processing and quality assurance of data they are not optimized for data retrieval. To overcome this deficiency data marts pull data from the transactional database and provide it in a format that makes querying, retrieving and reporting on data easy and efficient. Additionally, the data mart can restrict the quality level of data pulled making it an ideal repository for "reportable" data, i.e. data that has met the quality standards of RRES-RS. Eventually a data warehouse will sit in between the transactional database and SMART data mart. The transactional database will "feed" the warehouse and the warehouse will feed the SMART data mart. This architecture will allow for the creation of other data marts to "feed" other applications in the future.

A datamart is constructed differently than a transactional database. It is composed primarily of a fact table and related dimension tables arranged in what are called schemas. There are two types of schemas: star and snowflake. A star schema has each dimension table linked to one fact table. A snowflake schema has one or more dimension tables joined to a main dimension table instead of to the

fact table. The SMART datamart uses a star schema. Star schemas are easier to use for data retrieval as they require fewer joins and are generally easier to manage.

Fact & Dimension Tables

RRES-R analytical data is designed around a chemical analysis result fact table. It contains one record for each laboratory analysis result. This amounts to nearly 5 million records. The fact table has indexes that contain keys to the dimension tables. The dimension tables contain the attributes of the fact records. The fact table doesn't contain descriptive information or any data other than the numerical results and the index fields that relate the facts to corresponding entries in the dimension tables. Basically, the only "descriptive" information in the fact table is the numerical results of the laboratory analysis.

The dimension tables contain descriptive attributes related to fact records in the fact table. Some of these attributes provide descriptive information; others are used to specify how fact table data should be summarized to provide useful information to the analyst. Dimension tables contain hierarchies of attributes that aid in summarization. Dimensional modeling produces dimension tables in which each table contains fact attributes that are independent of those in other dimensions. Queries use attributes in dimensions to specify a view into the fact information. Subsequent queries might drill down along one or more dimensions to examine more detailed data. Examples of dimension tables include analytical suite and method, sampling depths, media, etc.

GIS Data

Storing attribute data in shapefiles or in SDE is an appropriate solution when that data is of reasonable size and is not maintained elsewhere. This is especially true if GIS data is coming from multiple source organizations and therefore is not under one ownership. If data needed for reporting is large, complex and of a relational nature, flat file attribute handling is not very efficient. The approach taken with SMART is to keep the minimum amount of data in the attribute table (usually just GIS oriented data) and instead retrieve data from the datamart using feature identifiers.

ArcIMS as a Data Retrieval Interface

The main feature of the SMART ArcIMS interface is sampling locations. One sampling location can have information about the location itself, the samples taken at the location and the results associated with the samples. To maintain all that information in the location layer would have to independently produce 3 million plus records. Even if managed in SDE this would require coordination of multiple databases and would not solve data retrieval speed issues. A simpler solution is to use the location ID to retrieve data in the datamart and then display that data with a reporting engine such as Crystal Reports or Microsoft office components. For example, in time tests, we were able to pull 500 pages worth of data over the web in less than 20 seconds. It would be impossible to achieve these speeds in a transactional data pull.

Specialized tools were developed in ArcIMS and methods were developed to allow the tabular based part of the application "talk" with the GIS interface. Specifically a tool was developed to allow users to select location point features and then retrieval all related sampling data associated with those locations. Again, we are talking about pulling back in many cases several hundred pages worth of data with multiple criteria selected. This had to occur in a timely fashion. Datamart design allows for

complex queries to be returned to the user in an acceptable timeframe. Additionally, complex queries performed on the tabular side needed to be reflected back to the GIS interface. Through use of the ArcIMS acetate layer users are able to view the locations of the results pulled back by their datamart queries. These tools allowed for the seamless integration of both datamart and GIS data as well as the tabular and GIS interfaces.

Conclusion

In brief, there are significant benefits to be accrued by using the methodology discussed in this paper. The benefits include, ease of management, data security and integrity, and speed of data retrieval. It has been demonstrated that this approach integrates well into a GIS framework and can be used successfully on the web.

Steven Scherma
reVision, Inc.
RRES-Environmental Characterization and Remediation
Risk Reduction and Environmental Stewardship Division
M992
Los Alamos National Laboratory
Los Alamos, NM 22102
Phone: 505.665.3532
Fax: 505.665.4747
E-mail: scherma@lanl.gov

Stephen Bolivar
Information Management Team Leader
RRES-Environmental Characterization and Remediation
Risk Reduction and Environmental Stewardship Division
M992
Los Alamos National Laboratory
Los Alamos, NM 22102
Phone: 505.667.1868
Fax: 505.665.4747
E-mail: bolivar@lanl.gov

Paper #: 2142

Internal:

ER ID#: ER2004-0407
LAUR #: